International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 93-98
e-ISSN:2278-621X

PREDICTION OF LEAFY COTYLEDON AND WUSCHEL PROTEIN USING SUPPORT VECTOR MACHINE

Kavyashree[1], Manasa[2] *and* Hemalatha N[3]

Abstract - Leafy cotyledon(LEC) protein is responsible for somatic embryogenesis and the initiation and maintenance of the embryonic pathway. Wuschel(WUS) is a member of WUSCHEL-related homeobox (WOX), a large group of transcription factors characteristically found in plants. It contains a conserved domain required for plant development by controlling cell division and differentiation. In this work we have used machine learning classification algorithm, Support Vector Machine(SVM) to create a predictive model for LEC and WUS. For creating the model for each protein, novel sequences of LEC and WUS were obtained from Genome Wide Analysis of these proteins in Palm plant. Independent data set and cross validation set were used to measure the performance of the predictive model. This was repeated for 3 kernel of SVM for the proteins.The best predictive model was obtained for C-terminal 25 composition using linear kernel with an Accuracy of 83% for LEC proteins. Similarly, for WUS proteins the predictive model was obtained for N-terminal 30 composition using linear kernel with an accuracy of 90%.
Keywords: Leafy Cotyledon, Wuschel, SVM.

I. INTRODUCTION

Leafy cotyledon1 (LEC) is a central regulator that control many aspects of embryogenesis and essential to induce embryo development in vegetative cells. It is also involved in inhibiting premature germination [1].The gene encodes a transcription factor, the CCAAT box-binding factor HAP3 subunit. LEC genes assembles only during seed development in embryo cells and in endosperm tissue. It induces the expression of embryo-specific genes and initiates generation of embryo-like structures. LEC activates the transcript of genes required for both embryo morphogenesis and cellular differentiation.

Somatic embryogenesis is a process where a plant or embryo is derived from single somatic cells. Somatic embryos are developed from plant cells that are not commonly involved in the development of embryos that is normal plant tissue [3].Homeobox genes encode for transcription factors containing a DNA binding domain called homeo domain with about 60 amino acids, which forms three helixes in space [4]. The homeobox gene was first identified in Drosophila [5,6]. Subsequently, more homeobox members have been stated in most eukaryotes [7]. WOX (WUSCHEL-related homeobox) is the member of ZIP superfamily belonging to homeobox proteins family [8]. Wuschel (WUS), a homeodomain protein previously catagorized as a key regulator for the specication of meristem cell fate.Wuschel protein (WUS), in accumulation to its role in controlling meristem development, also plays a critical role in the maintenance of embryonic cell identity. WUS expression induced somatic embryogenesis, suggestive that WUS promotes embryonic identity. It is a transcription factor that plays a central role throughout early embryogenesis, flowering and oogenesis , probably by regulating expression of specific genes.

[1] *Aloysius Institute of Management and Information technology, Mangalore,Karnataka , India.*
[2] *Aloysius Institute of Management and Information technology, Mangalore,Karnataka , India.*
[3] *Aloysius Institute of Management and Information technology, Mangalore,Karnataka , India.*

In this study we have used novel sequences of LEC, WUS which was obtained from Genome wide analysis. The machine learning statistical classifier SVM was used to create the predictive model for the proteins. The prediction accuracy of the SVM based classifier was assessed by two distinct approaches: cross validation test and independent dataset tests.For both the test we have used 3 composition approaches which are discussed in the result section.

II.METHODS

A. Dataset-

The Dataset used in this study consist of 30 LEC and 20 WUS proteins which was obtained from genome wide analysis in Palm plant. Out of the novel proteins 20 LEC, 16 WUS and were taken as positive training set and rest of each protein were used for the test set. Non LEC and non WUS from palm plant were taken as negative training and test set.

B. SVM-

In this study, predictions with classification method was evaluated using a strong machine learning technique, SVM (Support vector machine).SVM was introduced in 1992, by Boser, Guyon, and Vapnik in COLT-92. SVM, a machine learning method, has been utilized for many kinds of classification, regression and pattern recognition problems. SVM is a supervised machine learning technology originatedhypothetically on statistical learning theory [10]. Here to implement SVM, SVM light package was used, which allows to choose number of parameters and kernels (eg: linear, Polynomial, and radial basis function).Linear kernel is often recommended for text classification. The selection of kernel is very important in SVM and is analogous to selectingdesign in artificial neural network. In this paper, machine was trained using three kernels: linear, Polynomial, and radial basis function.

C. Features-

In order to produce different features based on amino acids, the frequency of occurrence of the 20 amino acids was considered. This method also created different standard window size among all the selected sequences i.e. N Terminal composition with 15, 20, 25 window size and C terminal composition with 15, 20, 30 window size. In another feature amino acid composition was also considered.

$$\text{Fraction of ith amino acid} = \frac{\text{Total number amino acid i}}{\text{Total number of amino acid in protein}}$$

D. Performance Evaluation-

To assess the performance of gene prediction tool, the standard prediction measures by Burset and Guigo were applied[11].Brief description of these parameters are given below:

i. Sensitivity: It gives the amount of correctly predicted proteins.
ii. Specificity: It gives the amount of in correctly predicted proteins.
iii. Accuracy: It gives the total number of predictions that were correct
iv. Precision: It is the proportion of the predicted positive cases that were correct.
v. F measure: It commonly used “average” of precision and sensitivity

These parameters can be calculated using following equations

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{FN}+\text{TN})} \quad (3)$$

$$\text{Accuarcy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \quad (5)$$

$$\text{F measure} = \frac{\text{Precision}*\text{Sensitivity}}{\text{Precision}+\text{Sensitivity}} \quad (6)$$

Where TP and TN are correctly predicted positive protein and negative protein, respectively. FP and FN are wrongly predicted genes respectively.

The Matthews correlation coefficient is used in machine learning as aamount of the quality of binary classifications, introduced by biochemist Brian W. Matthews in 1975[9]. It takes into account positives as true and negatives as false and is generally regarded as a stable measure which can be used even if the classes are of very varying sizes. The MCC is in principle a correlation coefficient between the experimental and predicted binary classifications; it proceeds a value between −1 and +1. A coefficient of +1 denotes a perfect prediction, 0 no better than random prediction and −1 denotes total disagreement between prediction and observation. Positive MCC value stands for better prediction performance. MCC is calculated using the equation 7.

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FN})}} \qquad (7)$$

For evaluating the performance of the prediction tool, independent data test and cross validation test were carried out. In independent data test validation, training dataset it is considered to be independent of one another. Cross validation is a model validation method for evaluating how the outcomes of a statistical analysis will simplify to an independent data set. In this work we have used 10 fold cross validation using SVM light.

III. RESULTS

Testing of SVM on independent dataset for LEC protein in palm plants achieved an accuracy of 83% (fig 1) with an MCC value of 0.68 using linear kernelfor C terminal 25 composition of the amino acid in the protein.

Similarly, for cross validation test for LEC protein in palm plants achieved an accuracy of 83% with an MCC value of 0.68 usinglinear kernel for C terminal 25 composition of the amino acid in the protein. (Table1).

Testing of SVM on independent dataset for WUS protein in palm plants achieved an accuracy of 90% (fig 2)with an MCC value of 0.82 using linear composition of the N terminal 30 composition of amino acid in the protein.

Similarly, for cross validation test for LEC protein in palm plants achieved an accuracy of 90% with an MCC value of 0.82 usinglinear kernel for N terminal 30 composition of the amino acid in the protein.protein (Table 2).

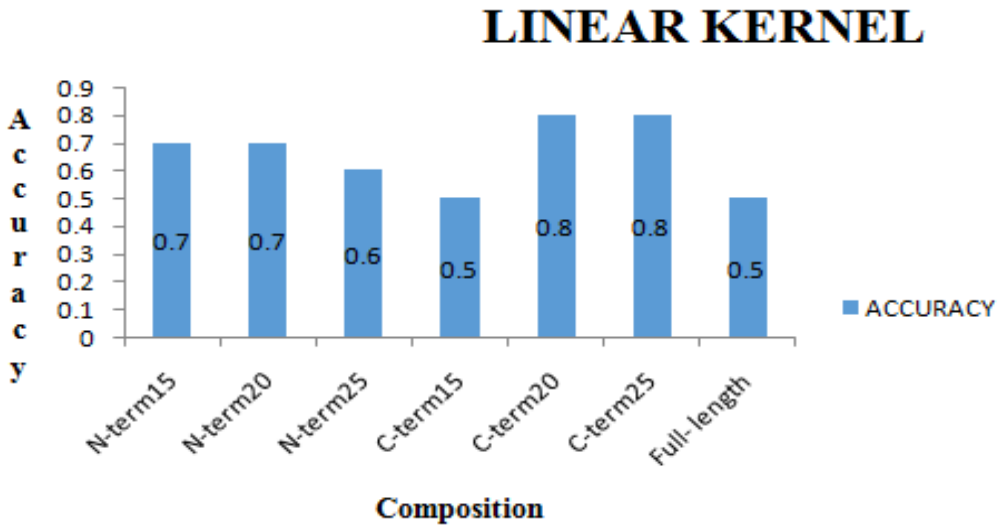

Figure 1. Performance Chart of Accuracy for different composition methods for linear kernel for the Independent and cross validation data set test for leafy cotyledon protein.

Table 1: Analysis of Independent Data set and Cross validation test for leafy cotyledon protein

		Independent						Cross validation					
Composition	Kernel	SN	SP	Precision	Accuracy	Fm	MCC	SN	SP	Precision	Accuracy	Fm	MCC
Nterm15	linear	0.73	0.6	0.65	0.67	0.35	0.34	0.73	0.6	0.65	0.67	0.34	0.33
	Poly	0.8	0.67	0.71	0.73	0.38	0.47	0.8	0.67	0.71	0.73	0.38	0.47
	rbf	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.3	0
Nterm20	linear	0.8	0.53	0.63	0.67	0.35	0.35	0.8	0.53	0.63	0.67	0.35	0.35
	Poly	0.8	0.4	0.57	0.6	0.33	0.22	0.8	0.4	0.57	0.6	0.33	0.22
	rbf	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
Nterm25	linear	0.93	0.27	0.56	0.6	0.35	0.27	0.93	0.27	0.56	0.6	0.35	0.27
	Poly	0.8	0.53	0.63	0.67	0.35	0.35	0.8	0.53	0.63	0.67	0.35	0.35
	rbf	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
Cterm15	linear	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
	Poly	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
	rbf	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
Cterm20	linear	0.87	0.8	0.81	0.83	0.42	0.67	0.87	0.8	0.81	0.83	0.42	0.67
	Poly	0.87	0.73	0.76	0.8	0.41	0.61	0.87	0.73	0.76	0.8	0.41	0.605
	rbf	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
Cterm25	linear	0.93	0.73	0.78	0.83	0.42	0.68	0.93	0.73	0.78	0.83	0.42	0.68
	poly	0.8	0.8	0.8	0.8	0.4	0.6	0.8	0.8	0.8	0.8	0.4	0.6
	rbf	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
	linear	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
	poly	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
Full length	rbf	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0

Table 2. Analysis of Independent Data set and Cross validation test for Wuschel protein.

		Independent						Cross validation					
Composition	Kernel	SN	SP	Prec	Acc	Fm	MCC	SN	SP	Prec	Acc	Fm	MCC
Nterm20	linear	0.69	0.88	0.85	0.78	0.58	0.57	0.69	0.88	0.85	0.78	0.58	0.57

	Poly	0.81	0.94	0.93	0.88	0.75	0.76	0.81	0.94	0.93	0.88	0.75	0.76
	rbf	0.125	1	1	0.56	0.13	0.25	0.12	1	1	0.56	0.12	0.25
Nterm30	linear	0.81	1	1	0.90	0.81	0.82	0.81	1	1	0.90	0.81	0.82
	Poly	0.75	0.93	0.92	0.84	0.69	0.69	0.75	0.93	0.92	0.84	0.69	0.69
	rbf	0.125	1	1	0.56	0.12	0.25	0.12	1	1	0.56	0.125	0.25
Nterm40	linear	0.75	0.87	0.85	0.81	0.64	0.62	0.75	0.87	0.85	0.81	0.64	0.62
	Poly	0.68	0.93	0.91	0.81	0.63	0.64	0.68	0.93	0.91	0.81	0.63	0.64
	rbf	0.125	1	1	0.56	0.12	0.25	0.125	1	1	0.56	0.12	0.25
Cterm15	linear	0.68	0.56	0.61	0.62	0.42	0.25	0.68	0.56	0.61	0.62	0.42	0.25
	Poly	0.68	0.68	0.68	0.68	0.47	0.37	0.68	0.69	0.69	0.69	0.47	0.37
	rbf	0.81	1	1	0.90	0.81	0.32	0.81	1	1	0.90	0.81	0.32
Cterm20	linear	0.75	0.37	0.54	0.56	0.40	0.13	0.75	0.37	0.54	0.56	0.40	0.13
	Poly	0.75	0.56	0.66	0.65	0.5	0.35	0.75	0.56	0.66	0.65	0.5	0.35
	rbf	0.18	1	1	0.59	0.18	0.32	0.18	1	1	0.59	0.1875	0.32
	linear	0.87	0.25	0.53	0.56	0.47	0.16	0.87	0.25	0.53	0.56	0.47	0.16
	poly	0.81	0.31	0.54	0.56	0.44	0.14	0.81	0.31	0.54	0.56	0.44	0.14
Cterm30	rbf	0.18	1	1	0.59	0.18	0.32	0.18	1	1	0.59	0.18	0.31
	linear	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
	poly	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0
Full length	rbf	1	0	0.5	0.5	0.33	0	1	0	0.5	0.5	0.33	0

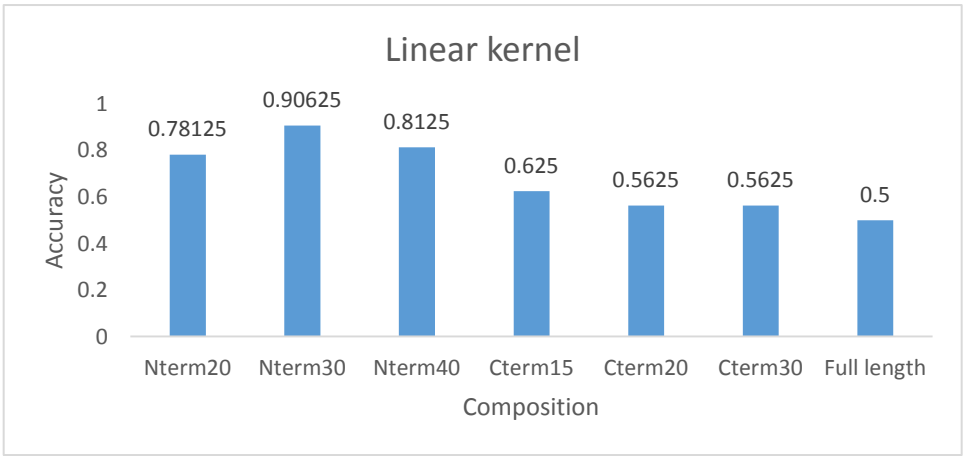

Figure 4. Performance Chart of Accuracy for different composition methods for linear kernel for the Independent and Cross Validation Data set test.

REFERENCES

[1] T. Lotan, M. Ohto, K.M. Yee, M.A. West, R. Lo, R.W. Kwong, K. Yamagishi, R.L. Fischer, R.B. Goldberg, J.J. Harada," Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells", Cell, pp.1195–1205, 1998.

[2] R.W. Kwong, A.Q. Bui, H.Lee, L.W.Kwong, R.L.Fischer, R.B.Goldberg,J.J.Harada ,"LEAFY COTYLEDON1-LIKE defines a class of regulators essential for embryo development", Plant Cell, pp. 5-18, 2003.

[3] Jianru Zuo, Niu. Qi-Wen, Frugis. Giovanna, Chua. Nam-Hai, "Somatic Embryogenesis in *Arabidopsis thaliana* Promoted by the WuschelHomeodomain Protein", Springer, pp. 279-281, 2003.

[4] C. Wolberger "Homeodomain interactions", Current Opinion *in Structural* Biology, pp.62-8, 1996.

[5] W.J.Gehring, M. Affolter, T. Burglin"*Homeodomain proteins", Annual Review of Biochemistry,pp.* 487-526,1994.

[6] *E. van der Graaff, T. Laux, S. A. Rensing,* "*The WUS homeobox-containing(WOX)proteinfamily",* Genome Biology pp.248,2009.

[7] R. Derelle, P. Lopez, H. L. Guyader, and M. Manuel,"Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes", Evolution and Development, pp.212-219,2007.

[8] G.Bharathan, B.J. Janssen, E.A.Kellogg, N.Sinha,"Did homeodomain proteins duplicate before the origin of angiosperms, fungi, and metazoa?" *Proceedings of the National Academy of Sciences of the United States of America*. pp.13749–13753,1997.

[9] B.**Wu**, T. **Abbott** , D.**Fishman**, W.**McMurray**, G.**Mor**, K.**Stone**, D.**Ward**, K .**Williams** , H.Zhao,"Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data",pp. 1636-1643, 2003.